DOCUMENT RESUME

ED 194 591 TM 800 721

AUTHOR Powell, J.C.
TITLE Wrong Answers on Multiple-Choice Achievement Tests: Blind Guesses or Systematic Choices?.
PUB DATE Apr 80
NOTE 52p.: An earlier version was presented at a joint meeting of the Psychometric Society and the American Association for Applied Mathematics (Hamilton, Ontario, August, 1978).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Elementary Education: Foreign Countries: *Goodness of Fit: *Guessing (Tests): *Mathematical Models: *Multiple Choice Tests: Regression (Statistics): Test Construction: Test Interpretation: Test Theory
IDENTIFIERS Canada: Proverbs Test

ABSTRACT

A multi-faceted model for the selection of answers for multiple-choice tests was developed from the findings of a series of exploratory studies. This model implies that answer selection should be curvilinear. A series of models were tested for fit using the chi square procedure. Data were collected from 359 elementary school students ages 9-12. For each student, the data included the frequency of selection of each of four alternatives on the first three items of Gorham's Proverbs Test. None of the linear models fitted, even when wrong answers were included in the model. The curved-line model which was developed from Bock's MULTIQUAL and which included quadratic and cubic polynomials was the only model to fit. Estimates for explained variance suggest that about 25 per cent may be available from right answers alone. More than twice as much may be available from a curved-line procedure. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194591

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Wrong Answers on Multiple-Choice Achievement Tests: Blind Guesses or Systematic Choices?

by

J. C. Powell
Faculty of Education
University of Windsor

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J.C. Powell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

© April 1980, J. C. Powell

An earlier version of this paper was presented at the joint meeting of the Psychometric Society and the American Association for Applied Mathematics in Hamilton, Ontario. August 1978.

TM 800 721

ABSTRACT

This study is the first in a series of three which are designed to validate or to refute current test theory. As a confirmatory study the methodological and the philosophy of science considerations are given more than the usual direct attention. A multi-facetted model for the selection of answers for multiple-choice tests was developed from the findings of a series of exploratory studies which have been completed prior to the present one. This model implies that answer selection should be curvi-linear. A series of models were tested for fit using the χ^2 procedure. None of the linear models fitted, even when wrong answers were included in the model. The curved-line model which was developed from Bock's (1973) MULTIQUAL and which included quadratic and cubic polynomials was the only model to fit. Estimates for explained variance suggest that about 25 per cent may be available from right answers alone. More than twice as much may be available from a curved-line procedure. The possibility that in addition to invalidating the "know-guess" hypothesis, that these findings might represent a breakthrough in measurement technology was discussed, and implications to measurement and to classroom instruction were drawn.

WRONG ANSWERS ON MULTIPLE-CHOICE ACHIEVEMENT TESTS: BLIND GUESSES OR SYSTEMATIC CHOICES?[1]

INTRODUCTION

This paper is the first in a series of three "confirmatory" studies which have been designed to validate or to refute the currently accepted theory for test construction.

The approach for this series has been to examine the distribution properties of learner responses on an answer-by-answer basis, including all of the wrong answers. Answers are treated as separate decision events and as such as nominal or categorical in nature.

The collective results of these three studies has been to cast serious doubts upon the validity of current practice for many educational applications. This particular study addresses the current practice of interpreting multiple-choice responses as being dichotomous (KNOW-GUESS) events.

Some background, both philosophical and historical, is needed to put this essentially methodological study into perspective.

a. Philosophy of Science Background.

With the advent of electronic data processing, a considerable increase in the sophistication of research procedures has become possible. As a result of these advances, in the physical sciences, a four step approach to research seems to have been emerging. These four steps are 1) explore a data set for a model, 2)confirm the viability of the model, 3)test the applicability of the model in a replication and 4)refine the model using inferences from it for further experimentation. The procedure can be made more powerful by developing more than one competing model to test for viability.

As Glaser and Strauss (1967) pointed out, social scientists

---

1. This paper was first presented to the joint meeting of the Psychometric Society and the American Association for Applied Mathematics, Hamilton, Ontario. August 1978. The large data set reported in the replication part of this study was collected with the financial assistance of the Ontario Educational Research Council.

have often attempted to conduct research at the level of the last of these four steps without having gone through the first three of them. They have begun with a theory derived from the original problem rather than beginning with the directly observable data.

The contention will be explored, in this paper, that this same procedural error may have been made by educators and educational researchers with respect to the determination of achievement status of learners in their intellectual development.

Current practice has arisen from a substantial and valid historical base. When human knowledge was preserved in, first oral and then hand-copied written tradition, the intact preservation of the essential parts of this tradition was of profound importance. The tradition of an absolute reality (of one "right" answer) and of a deterministic reality has been so substantial that even in the physical sciences the realization of the fallacy of this position is hardly more than 100 years old. In fact, this distinction between a deterministic and a probabilistic reality has been the heart of the phenomenological fallacy which has distinguished 19th Century Science from 20th Century science in all areas of knowledge.

Within the context of education, determinism seems to have been tacitly retained in the concept of the pre-determined outcome expectation as exemplified by the exclusive use of "right answers" as a basis for the determination of achievement status. If this source for learner appraisal can be demonstrated to be the only available source of viable data, then its use as the sole determiner of success will have been validated.

From a philosophy of science position, then, all of the conditions for a theoretical break-through seem to be present. The current approach seems to be a theory which is built upon assumptions which may not possess prior observational validity. There is in the same problem domain a set of direct data (wrong answers) which are commonly ignored. Finally, the justification for the use of the data used, and the ignoring of the data which is overlooked seems to arise from a deterministic orientation; a presumption about reality which has proven to be incorrect in many other fields of science.

In this last comment, methodological and philosophical issues

conjoin to provide the theoretical basis for the problem in hand. In order to establish the issues with clarity, it would probably be helpful to recall the fundamental assumptions of current test theory. There are three of these assumptions. Each of them and their mathematical derivatives are presented in Figure 1.

FIGURE 1

PRINCIPAL ASSUMPTIONS OF
CURRENT TEST THEORY

Behavioral Assumption	Mathematical Equivalent
1. THE KNOW-GUESS HYPOTHESIS The learner either knows the answer or guesses (blindly).	1. $X_{ij}=\partial$; i.e. the answer given by person i on item j is a 1 for a right answer and a 0 for a wrong answer.
2. THE CONTINUOUS LEARNING HYPOTHESIS Learning is continuous and cumulative. (Graphically learning would appear as a straight line when age and achievement are equated).	2. $X_i=\sum_{j}^{n} X_{ij}$; i.e. the score of a person on a test is the sum of all the 1's and 0's from the items on the test.
3. THE LINEAR MEASURABILITY HYPOTHESIS Achievement is linearly measurable. (Departures from the straight line in # 2 are the product of meaningless measurement error and other random events.)	3. $X_i=T_i+E_i$; i.e. the observed score (X_i) is a linear combination (sum) of the person's linear true score (T_i) and an error of measurement (E_i).

In behavioral terms, the most primitive of these assumptions proposes that in the making of an answer selection the learner either knows the correct answer or guesses (perhaps blindly) when s/he is responding to a multiple-choice test. This assumption leads directly to a procedure which has the effect of treating all other answers as a more or less homogeneous category of behavior best described as "lack of knowledge." The mathematical conversion of this assumption into its binary equivalent when the answer vector is being scored, wherein all "right answers" as predetermined from the scoring key are made a one (1) and all other answers are reduced

to a zero (0), completes the homogenization of this second category of behavior. This procedure would be valid, provided that either the behavioral phenomenon being reduced in this manner is not systematic and did not, therefore, discriminate among learners, or provided that these discrimination proved to be irrelevant to the measurement of intellectual development. As a result, before any direct attack upon the validity of any of these three assumptions can be mounted, both the statistical and the behavioral relevance must be demonstrated.

This last comment leads directly to the historical background of this problem.

b. Historical Background.

In the case of the present author, this present problem arose from pragmatic considerations in the day-to-day operation of a high-school classroom. It was observed during item analysis of tests that certain wrong answers seemed to reflect information which was useful to the teacher in a diagnostic sense. In this case, greater precision concerning the instructional needs of these learners could be obtained from the specific choice of a particular wrong answer than was available from the mere fact that these and other individuals had answered this question incorrectly. In this early stage, a clear distinction could be drawn between the determination of a learner's instructional needs (diagnosis) and the assessment of his or her achievement status (evaluation.) However, in the persuit of the practical issues of trying to improve the quality of wrong answer information in order to make better use of it, the distinction between these two poles (diagnosis and evaluation) seemed to fade.

In effect, the present author was engaging in the exploration of the distribution properties of a behavioral phenomenon (wrong answers) long before the scientific implications of this approach had been realized or the fact that these properties were being largely overlooked by other researchers had been recognized.

Such work as has been done with wrong answers on multiple-choice tests seems to fall into five broad areas. The most exten-

sively explored issue has been that of formula scoring of "correction-for guessing." The second, actually an off-shoot of the first, has involved attempts at the development of procedures for the consideration of "partial knowledge" in the scoring process. The third has involved attempts to obtain diagnostic information from wrong answers. A few researchers have considered the developmental properties of these selections, and an occasional look at their information properties can also be found in the literature. The first three of these clearly maintain the tacit assumption that there is, of necessity, a single right answer to each problem. The other two areas are less clear in their stance upon this issue.

Formula scoring was originally an attempt to come closer to a learner's true score by removing from his or her observed score that proportion of answers which might reasonably be attributed to "lucky guesses." Results of these attempts have met with mixed success, with some researchers finding improved reliability (Jacobs and Vandenventer 1967) and others not (Kansup and Hakstain 1975). Such mixed results can, theoretically, occur for a variety of possible reasons. One is that wrong answers are generally not related to right answers. There is an inner contradiction to this position since wrong answers, collectively, are linearly dependent in relation to right answers and, therefore, should correlate at the level of $r = -1.00$ rather than $r = 0.00$. The more likely possibility is for wrong answers to be related to right ones in an interactive or curvilinear fashion. This possibility would explain the mixed results, but has the problem that if it is true, then current test theory, which assumes that the only meaningful relationships within these data involve straight lines, could be invalid.

The procedures for accommodating partial information (Shuford, Albert and Massengill 1966) have been shown to add a "risk taking" dimension to the measurements, (Hakstain and Kansup 1975) but the connection with willingness to risk and "knowledge" in a deterministic sense is unclear at best. However, it has long been known that partial information may not dispose a person to come closer to the "right" answer. Over-simplification is a logical fallacy of Greek antiquity. The "partial information" approach does not consider

the possibility that insufficient information could have the same effect as "misinformation" of predisposing the learner to select, systematically, particular wrong answers. Nor does this approach consider the logical possibility the these two knowledge states might predispose the selection of different wrong answers, thus producing effective discrimination among learners from the wrong alternatives.

This last comment leads directly into the concept of using wrong answer information for diagnostic purposes. Very little is in the literature on this issue, although such use of this information is common practice among teachers who use the "whole child" approach to teaching. Instead, as is evident in the competency testing movement, test designers are more likely to produce single skill items and to use the "right-wrong" nature of the responses to provide the diagnosis. Imbedding more than one skill into a single item and then using the nature of the error to provide the diagnosis is not a common practice. For this reason, outcome referenced and other diagnostic tests tend to be long and cumbersome. This inclination may be inferred as representing the tacit acceptance of the deterministic position as well.

The most certain example of a developmental interpretation of wrong answer data is found in the administration manual for the pre-school version of Raven's Progressive Matrices (Foulds and Forbes 1965) In this case, these authors present a table which classifies particular errors by type and by age level of typical selection. They suggest that developmental placement can be made with increased certainty if this information were used. On the issue of determinism in the underlying assumptions, these authors seem to imply a single developmental pathway followed by all learners. A probabilistic model would be more likely to propose multiple pathways, none of which were precisely fixed or irrevocable.

Conversations at conferences have led to several suggestions that particular individuals may have explored the information properties of wrong answers, but these studies, in finished form seem to be remarkably difficult to obtain. The best example in the literature seems to be the nominal classification attempt by Bock (1972) in

which he looked at a number of vocabulary items which were administered to college freshmen. Using an information approach, he concluded that no gains would be achieved for individuals above the median. This finding would seem to support the diagnostic use of wrong answer information. On the other hand, vocabulary items might conceivably be construed as being answerable upon a "know-guess" basis more than other types of multiple-choice item. The dichotomous model should fit these items, perhaps, better than others.

In summary, the literature concerning the instructional use of wrong answer information seems to be both sparce and contradictory. Nearly all of it seems to treat the wrong answer as if it were a screen or a filter to the right answer, rather than as a source of precise information in its own right. No direct challenge to the deterministic position, and hence to the "know-guess" hypothesis seems to have been contemplated by many of these researchers, nor have the procedures used made such a challenge directly possible.

c. Findings from Exploratory Studies.

As mentioned above, before any direct challenge to current practice can be made, both the statistical and the behavioral relevance of wrong answer data must be established.

The present author had conducted several exploratory studies prior to this present confirmatory one (Powell 1968. 1970. 1976. 1977, Powell and Isbister 1974.) Of these, the APA paper (1976) summarizes all of the others.

The diagnostic value of wrong answers had been known for several years before a change of procedure led to a critically important observation. In the first of two Dissertation pilot studies, it was decided to change from mathematical and/or scientific content to language based content and to collect self-reports of the reasoning involved rather than using the internal structure of the item as a basis for deciding the nature of the error. In addition to finding that error types could be classified consistantly on the basis of the reasoning reports and that this classification replicated to another group, it was discovered that some people "over-read" the items and selected answers upon the basis of valid divergent interpretations

of these items, making their answers logically correct. (Powell 1968) This observation provided empirical support to Hoffman's (1962) criticisms of multiple-choice testing. More important for present purposes, the discovery that non-keyed answers could be logically correct by a limited number of testees implied a potential discrimination of achievement status which invalided the key to the test. Of course, it could also be construed as invalidating those particular items, but to remove these items also would remove the achievement discrimination this observation seems to make available. In other words, this observation provided a probabilistic result from a deterministic procedure. This implication presented the possibility that the issue was far more fundamental that the quality of a few items. To satisfy this dilemma, it would be necessary to show that alternate readings was a far more common phenomenon than was typically supposed.

To resolve this issue, a similar study using the same test with children (Grades 3 to 8 inclusive) rather than adults, once again collecting reasoning (this time using trained interviewers) was undertaken. The results of this study are also in the literature (Powell 1976, 1977.)

The best ordering (simplex) for the correlation matrix in this study led to a perfect marginal recapitulation of the age order. This observation was considered to provide clear evidence for a strong developmental property among these data. The reported reasoning led to the classification of wrong answers subsets which was internally consistant as based upon the logic of these reports for more than 50 percent of those who chose answers within a subgrouping of alternatives. The similar result for adults had been 64 per cent. Perhaps wrong answers stability increases with age.

The general properties of the reasoning reports produced two independent ordered categorical scales. One of them was clearly developmental, reflecting some of Piaget's findings and was related to the manner in which the items were interpreted. The second ordering was related to the strength of influence a particular aspect of the solution approach had upon its outcome. Item 18 as given in Figure 2 was chosen to illustrate these properties because it contains the interactive outcomes for both orderings.

FIGURE 2

WRONG ANSWERS MAY NOT BE GUESSES

PROVERB:

QUICKLY COME, QUICKLY GO.
(Easy Come, Easy Go)*

TRANSLATIONS:

a. ALWAYS COMING AND GOING AND NEVER SATISFIED.
Characteristic of 13 year olds.
"You Should Stick To A Job Til It's Finished."

b. WHAT YOU GET EASILY DOES NOT MEAN MUCH TO YOU.
Characteristic of adults.

c. ALWAYS DO THINGS ON TIME.
Characteristic of 8 year olds.
"That's What Teacher Always Says."

d. MOST PEOPLE DO AS THEY PLEASE AND GO AS THEY PLEASE.
Characteristic of 10 year olds.
"It Talks About Coming and Going."

*Item No. 18 from Gorham, Donald R. Proverbs Test, Psychological Test Specialists, 1956. Reproduced with permission.

The three developmental levels seem to be based upon the point-of-view taken by the testee and were inferred, based upon the reported reasonig to be: 1)a personalized interpretation, 2) an interpretation based upon literal meanings and 3) an interpretation based upon simple causal relationships.

The 8 year-old seems to be construing the question of the meaning of "quickly come, quickly go" in terms of his or her own physical movement about the Grade 3 classroom. Part of being quick "as the teacher always says" is getting things done on time. Since personal physical motion is a quite legitimate (though not proverbial) interpretation of the phrase "quickly come, quickly go. " and the inferential linkage to timing is both empirically and logically correct, this answer is valid, even though it is considered to be wrong from the arbitrary keying system. Admittedly, the keying is based upon a knowledge of the culturally ascribed meaning to this proverb. In this sense, this answer represents a "lack of knowledge" but the

selection of it, if the reported reasoning can be taken to be valid upon the basis of more than 50 per cent consistancy, can not be meaningfully construed as a "blind guess." Nor can the proposition that this selection contains no meaningful data concerning learner status in intellectual development be sustained.

The second level, searching for literal interpretations, seems characteristic of the reported 10 year-old. In this case, the logic of the fit is less sound. The implication was taken, since some of the learners admitted to this short-coming, that an inappropriate solution strategy was selected in this case. Thus the second of the three potency elements, solution strategy selected, is also evident in this selection of answer. The relationship between strategy selection and outcome seemed to be $\rho = .86$ as compared with the potency of 1.00 found for the developmental perspective. This latter contrasts with the $r = .50$ which is common for educational studies using total-correct scores as the data base. From these observations it may be inferred that the information base used to answer questions is of the weakest potency of these three elements in accounting for the selection of particular answers.

The third developmental level seems to be illustrated by the typical 13 year-old who seems to be taking the "never satisfied" part of the selected answer in a linear causal relationship to the proverb. Level four, of course is the ascribed answer. In a few cases, however, those who chose alternative A went beyond an internal explanation to say that they rejected B since the "never satisfied" implied essentially the ascribed meaning, and that since this proverb was usually used as a criticism, that A "went beyond" B in meaning. Perhaps this reasoning represents the "over-reading" already mentioned and may represent a fifth level of intellectual development.

Two important ideas may be inferred from this last observation. First, it may be possible for a particular answer to be selected by different individuals for differing reasons. In this case, a simple count of the frequency of any particular answer or answer type, with-determination of the reasoning behind that selection will probably lead to invalid conclusions about the achievement status of the learner. Perhaps, the relations among answers selection, already known

to be strongly developmentally based could separate people who chose any particular answer for different reasons. The second implication would seem to be that beyond a certain point, intellectual development may lead to divergency of approach at such a profound level that tests which require convergence to an ascribed answer put these learners at a disadvantage. This is essentially Hoffman's criticism. The inference by this time, however, had been supported using two different tests, and four different age groups. Since replication is a powerful tool for the development of evidence, these behavioral observations leave the practice of the binary interpretation of answer selection on very shakey behavioral grounds.

Using wrong answer sub-groupings as dummy variables in multiple-regression equations, wrong answers not only added to the variance, but also they came out earlier in the step-wise solutions, and accounted for more variance than did the right answers. This observation occurred using two different tests and two different age groups. In both of these studies, the dependent variables were total-correct scores upon independent measures of achievement. In both cases also the total variance accounted for rose to the 30 to 50 per cent range. Curved - line regressions seemed to do even better. From these observation, it seems reasonable to conclude that wrong answer data might be expected to contribute meaningfully to the assessment of intellectual achievement.

In summation, it would seem, from these exploratory studies, that the dichotomization of achievement into "knowledge" and "lack of knowledge" categories may constitute an invalid over-simplificaion of these data, particularly for test items where more than the lower level skills of recognition and/or recall seem to be involved. The studies may provide a grounded basis for the development of alternative behavioral models for answers selection which can be used to test the validity, or lack thereof, of current practice in educational testing.

ALTERNATIVE BEHAVIORAL MODELS

There are several possible general models which can be contrasted in order to explain answer selection. For the sake of convenience and simplicity, two such models will be considered in this present

study. One of these two will be deterministic in outlook, and the other will be probabilistic.

It will be easier to conceptualize the differences between these two if one is considered in terms of the impact of inheritance and the other in terms of the impact of environment, realizing that the actual situation probably lies somewhere between these two extremes.

In the "inheritance" model, cognitive development is considered to be determined at time of conception by the production of a new person who is under the developmental control of two interacting predispositions, temperment and intelligence. The tempermental aspect of the individual might be considered to be a normally distributed inclination along an expressive - reflective dimension, and intelligence might be considered to be normally distributed along a dull - bright dimension. Given this initial starting-point on a two dimensional matrix, development would be assumed to be linear and cumulative. In this case, the diversity observed among humans would be a mixture of starting-point and maturation time. In such a model, the best an educator can do for a learner would be to make the appropriate experiences available to a learner when s/he is ready to learn from them. The expectation would be that any cross-section of the population at any time would display a normal-curve distribution, with a similar pattern among random sub-populations. This model, with some refinements to include a low-potency environmental factor is essentially the tacit model which underlies current test theory. It is not necessary to belabor the way in which such concepts as norms, behavioral objectives, marking-to-the-curve and readiness can be derived from this model. The order of potency would seem to be 1) intelligence, 2) temperment and 3) experience within this model. It would seem to have considerable empirical support from the great volume of research which has been conducted over the past century. However, the validity of this model may be contingent upon the validity of the use of total-correct scores as the data-base. If the use of total-correct scores for achievement assessment is invalidated, as, indeed, this series of studies demonstrates, then all of these conclusions would need to be reexamined from at least one alternative perspective.

The easiset way to generate a contrasting model is to reverse the order of potency of these three variables, putting experience first and intelligence last in the sequence. Development is then assumed to be interactive rather than linear, and the pathways contingent rather than genetically determined. This model would also provide explanation for diversity and for the apparent normal distribution on totals of scores. However, the exploration of subgroup properties should display homogeneous clustering rather than continuous progress. Also there may be evidence of stasis in development and possibly even regression to earlier states. The clustering would be inferred from the interaction property and the tendancy of people in contiguous environments to have many experiences in common.

In this model, it would be expected that an educator could be an extremely potent force in an environment, in so far as current status can be known and contingencies manipulated.

In the narrowest possible interpretation of the "inheritance" model, there would be two possible reasons for the choice of a "right answer", either the learner KNEW the answer of made a LUCKY GUESS. All wrong answer selections will, of course, be GUESSES.

For the "environment" model, it would be necessary to invoke some form of contingency or decision model to explain answer selection. Figure 3 presents one possibility in the form of a multi-branch decision tree. It should be noted that this model contains five possible reasons for the selection of a right answer and four for a wrong one. This contrasting model, can be inferred from the "environmental" model, but it can also be inferred from the results of the exploritary studies already discussed.

INSERT FIGURE 3 ABOUT HERE

The relative potency of these various alternatives seems to be in the order given. Of special concern to the educator, is the inferred inability to distinguish between concepts which have been "memorized" and concepts which have been "understood." The apparent contingent nature of answer selection raises the hope that relation-

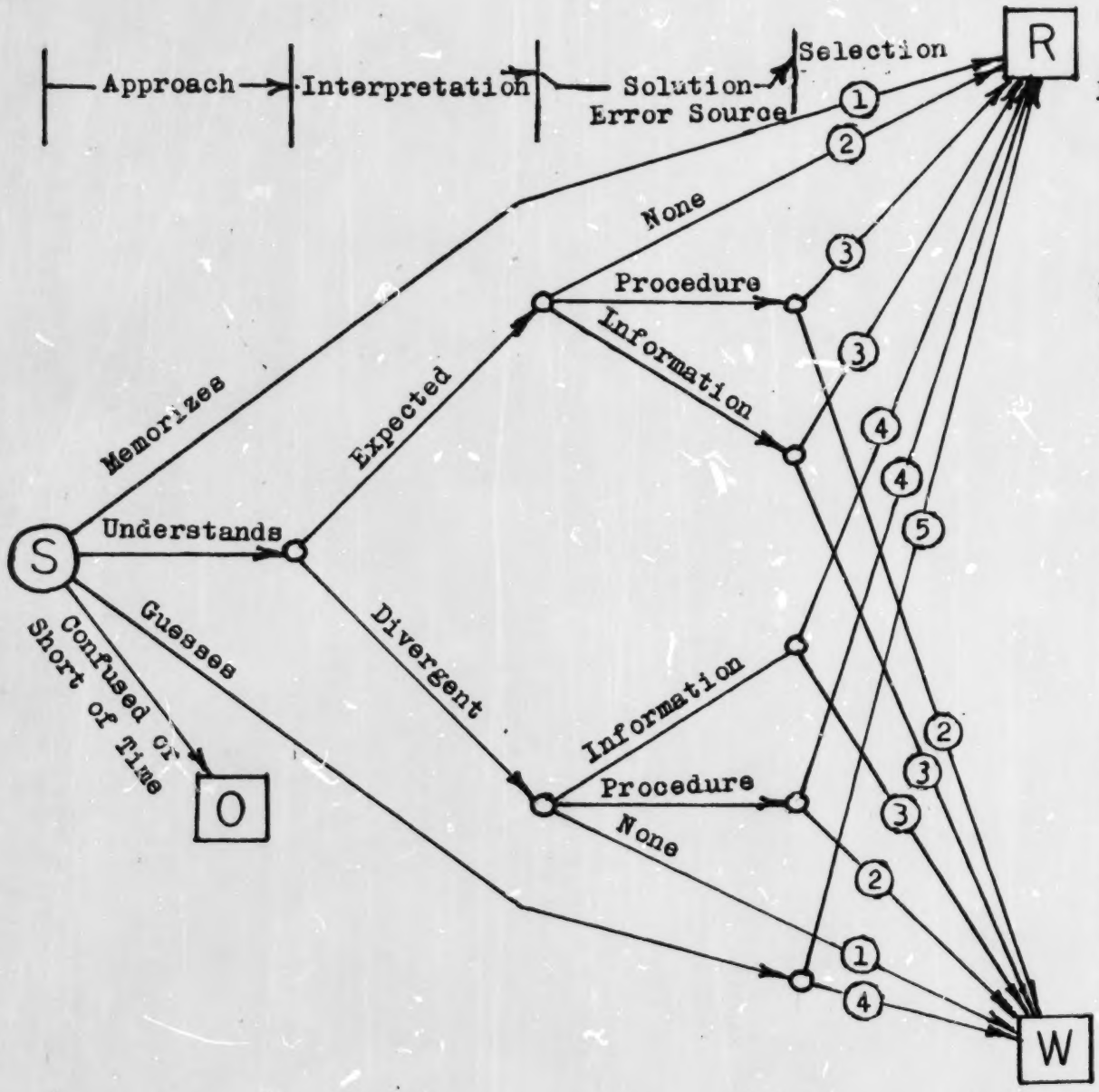


FIGURE 3

DECISION TREE ANALYSIS OF ANSWER SELECTION SHOWING STEPS IN THE DECISION PROCESS

KEY to the Reasons for the Selection of a:

Right Answer.
(In order of potency)

1. Blind memorization.
2. Understanding.
3. Mutually cancelling errors.
4. Errors in divergent interpretation.
5. Lucky guess.

Wrong Answer.
(In order of potency)

1. Valid divergent interpretation.
2. Procedural errors.
3. Information errors.
4. Unlucky guess.

Omitted Answer.

1. Listed on tree.

ships among specific answer selections may be used to make these distinctions.

There is little doubt that the best model will not be an either/or of these two. It is already known that total-correct scores generally account for about 25 per cent of the variance, with higher potency in highly contrived situations. In this context, it would be more likely to find current test theory to represent a special case of a more powerful and general model, in much the same way as Newtonian Mechanics seems to be a special case of the higher order models currently under development in physics.

At this point in this discussion, a demonstration that the dichotomization of the learner's response vector using the "know-guess" hypothesis may be invalid, is probably more nearly an approach to explaining why the current procedures tend of account for only about one quarter of the data, rather than to render current practice totally useless and obsolete. The problem, then in the study report which follows is to test the validity of this dichotomization of the test data as the only viable procedure for the assessment of school achievement.

THE PROBLEM

In a confirmatory study of this nature, it is necessary to test a series of mathematical models against a single data set. From the exploratory studies and from the alternative model for answer selection, it is reasonable to infer that the proportion of non-systematic events should be much lower in the second case than in the first.

For the "know-guess" model to be sufficient, the regression line for the right answer set of points, should account for the greatest bulk of the variance in a data set which includes the observed distributions for all alternatives. For the multi-interactive model to be successful, a curved line model must fit this data set and the regression line must fail. At some point between these two there should be a sudden drop in the value of χ^2 and the models beyond this point should fit, and those below it not. The location of this transition point in a chain of models should indicate the degree to

which dichotomization is appropriate for this particular data set.

Since research reports which involve the fitting of models to a data set are not very common in educational literature, the procedures and the results will be discussed in more than the usual detail.

PROCEDUEAL CONSIDERATIONS

In the usual procedure for statistical analysis a comparison is made between two distribution parameters of the same type, such as the means of two different groups. If these means are far enough apart, then the sources of the two groups can be assumed to come from different populations. The distance apart is determined from the known distribution of the means of distrubutions. A probability statement of the likelihood that means this far apart could be from the same population is stated.

When models are being tested, the nearness rather than the distance apart is the criterion for success. As a result, instead of seeking a probability of similarity of p = .05 or smaller to be able to infer difference, the model testing procedure is seeking a probability of similarity of p = .05 or greater in order to be able to infer sameness.

In the procedure used in this paper, a data set of 48 points was the basic observation. This set was composed of the frequency of selection of each of four alternatives for four age levels on three items. These were the first three items on The Proverbs Test (Gorham 1956) as collected in 1975. (Powell 1977) A mathematical model consisting of polynomials up to the third degree was generated using Bock's (1973) MULTIQUAL. This model was the curved-line model used in the study. The linear models were systematic patterns of proportions as will be described later. The output for the curved-line model in the form of proportions was also used. Only Item 1 will be reported, although similar results were found with all three items.

Finally, the proportions were converted to expected frequencies using the group sizes indicated in Table 1 and χ^2 statistics for the goodness of fit were calculated for all models in the chain.

The nearest whole number for the expected frequencies was used to simplify the hand calculations. After giving the details of the population used, the results for each model will be presented and discussed in turn.

THE POPULATION

A voluntary sample (in the sense that parental consent was given) of children in Grades 3 to 8 inclusive was collected in three suburban schools in May of 1975. There were approximately 90 children in each grade; 30 high, middle and low performers (N = 550).

For this present study only those children from the ages of nine to 12 inclusive were used. The 8 year-olds and the 14+ age group were dropped because of a group size of less than 60. The thirteen year-olds were dropped because of several indications that this group was unusual. One of these indications was the number of children who seemed to have "over-read" a number of questions. It was thought that the presence of such an unusual group might provide a spurious rejection of the linear model, hence its exclusion.

The age group frequencies for Item 1 as used in the current study are given in Table 1.

TABLE 1

Distribution of Subjects on Item 1 by Age

Age in Years	Frequency
9	100
10	86
11	99
12	74
Total	359

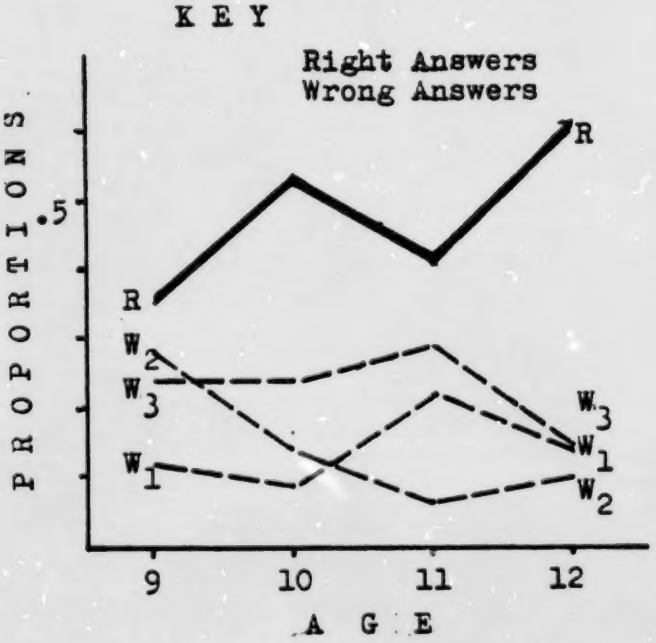
The full scope of this study is presented elsewhere (Powell 1976, 1977, 1979 and 1980).

Since these models are being tested upon part of the data used in the exploratory studies, it should be evident that the present

FIGURE 4

THE PROVERBS TEST

ITEM 1 OBSERVED DATA



Observed Proportions

Age	R	W_1	W_2	W_3
9	.36	.12	.28	.24
10	.53	.09	.14	.24
11	.42	.22	.07	.29
12	.61	.14	.10	.15
Ave.	.48	.14	.15	.23

Observed Frequencies

Age	R	W_1	W_2	W_3	Totals
9	36	12	28	24	100
10	45	8	12	21	86
11	41	22	7	29	99
12	45	10	8	11	74
Totals	167	52	55	85	359

procedure conforms to the established requirements for a "confirmatory" study.

RESULTS AND DISCUSSION

The basis data being used in this study are the frequencies of selection of each alternative for each age level. Figure 4 gives these observations:

INSERT FIGURE 4 ABOUT HERE

In Figure 4 the observed data are given in three different forms. First, they are represented graphically, using the proportions of selection for each cell. Second, the raw data are given in the table to the right below the graph. Third, the proportions of selection are given to the left below the graph. All of the models tested for fit in this study, of course, use these data for the observed values.

This format of presentation, in which both graphical and numerical presentations are given together, will be employed throughout this study in order to add clarity to the report.

MODEL ONE: All Answers "Blind Guesses"

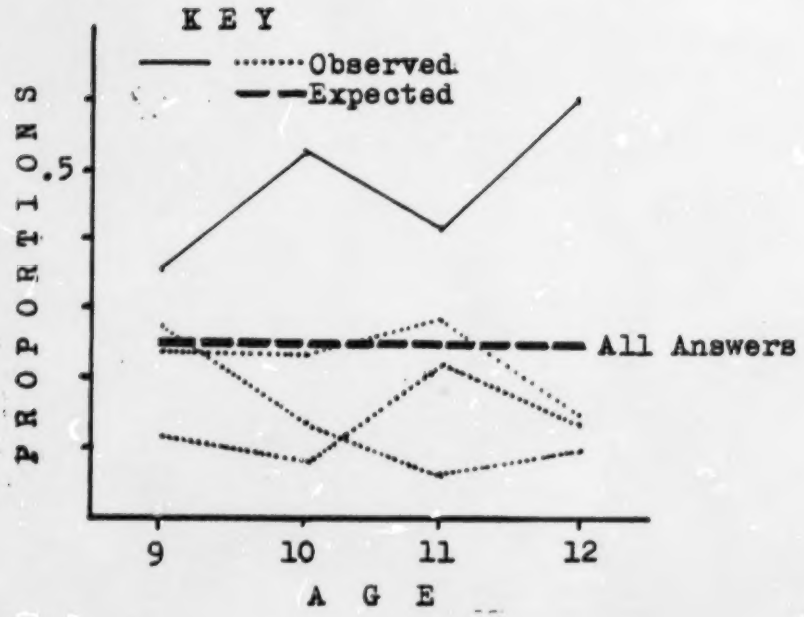
The issue under consideration in this study is whether or not a straight-line simplification of the proportions of selection of right answers alone, or right answers in linear combination with wrong answers, or in curvi-linear combination with wrong answers best fits these observed data. A secondary question involves the point in the sequence of models where failure changes to fit. Underlying this issue is the validity if the dichotomization of answers into "right" and "wrong" categories in an arbitrary manner before the test is administered.

Since the tested hypotheses involve the relative validity of a "know-guess" behavioral model in contrast to a more complex decision model approach to these data, the logical starting point

FIGURE 5

MODEL ONE: ALL ANSWERS CONSIDERED RANDOM

FAILS



MODEL ONE

Expected Proportions

Age	R	W_1	W_2	W_3
9	.25	.25	.25	.25
10	.25	.25	.25	.25
11	.25	.25	.25	.25
12	.25	.25	.25	.25

Observed/Expected Frequencies

Age	R	W_1	W_2	W_3	Totals
9	36_{25}	12_{25}	28_{25}	24_{25}	100
10	45_{22}	8_{21}	12_{22}	21_{21}	86
11	41_{25}	22_{25}	7_{24}	29_{25}	99
12	45_{18}	10_{19}	8_{18}	11_{19}	74
					359

χ^2 Table (d f = 11)

Age	R	W_1	W_2	W_3	Totals
9	4.84	6.76	.36	.04	12.00
10	24.05	8.05	4.55	.00	36.65
11	10.24	.36	12.04	.64	23.28
12	40.50	8.10	5.56	3.37	57.53
Totals	79.63 (7.82)	23.27	22.51	4.05	129.46 (19.68)

would seem to be to assume that every answer is a guess. In this case, the expected selection proportion, on a four alternative test, would be p = .25 for every alternative. The resulting value for χ^2 would be analogous to the available variance, since its size is indicative of the degree these data depart from purely random events.

Figure 5 presents the Model along with the χ^2 test for the "goodness of fit" of this model to the observed data.

INSERT FIGURE 5 ABOUT HERE

In Figure 5 are presented; 1) the graphs of the observations, 2) the graph of Model One, 3) the mathematical version of the model expressed in terms of proportions, 4) the frequencies used for the test of the model and 5) the χ^2 table. This pattern for presentation will be used for all seven of the models tested for fit.

As was expected, the test for the fit of this model produced a large χ^2 in the attempt. The comparative critical value for χ^2 at the p = .05 level is given in brackets below the obtained value.

In a later discussion this obtained value will be used as an estimate for the available variance, since it is directly related to this variable. The MULTIQUAL procedure, in the version employed, did not yet contain such a refinement.

It should be noted that once the expected frequencies have been obtained from the logic of the model, the standard χ^2 procedure is employed. The only unusual feature of this procedure is the interpretation of the probability level. In this case, of course, the expectations from Model One FAIL to fit the observations.

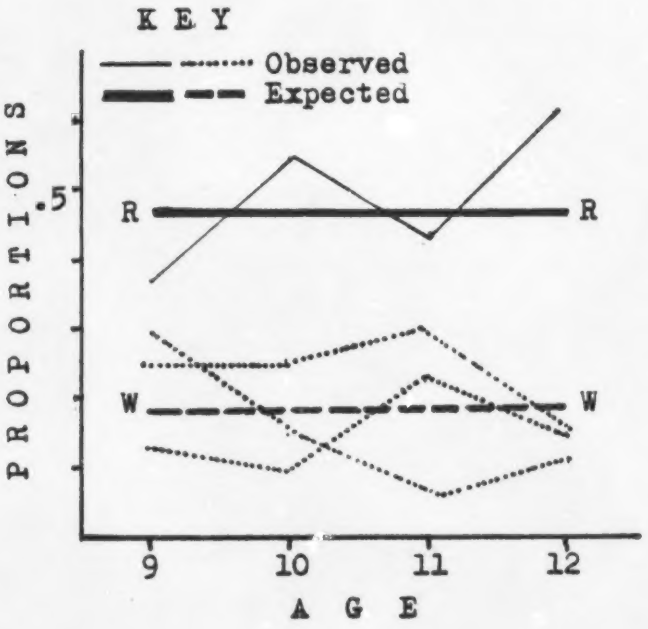
MODEL TWO: Average of Right Answers Meaningful

The logic of Model Two involves the assumption that only the right answers are meaningful. In this model no age trend is assumed. For this reason the right answers are graphed as a horizontal line. In Item 1, the right answer selection ratio is .46 (167 + 359 = .46). As a result the expected value for the wrorng answer ratios would be .18 ((1.00 - .46) + 3 = .18). The fact that this value is the

FIGURE 6

MODEL TWO: OVERALL
AVERAGE OF RIGHT ANSWERS

FAILS



MODEL TWO

Expected Proportions

Age	R	W_1	W_2	W_3
9	.46	.18	.18	.18
10	.46	.18	.18	.18
11	.46	.18	.18	.18
12	.46	.18	.18	.18

Observed/Expected Frequencies

Age	R	W_1	W_2	W_3	Totals
9	36_{46}	12_{18}	28_{18}	24_{18}	100
10	45_{41}	8_{15}	12_{15}	21_{15}	86
11	41_{45}	22_{18}	7_{18}	29_{18}	99
12	45_{35}	10_{13}	8_{13}	11_{13}	74
					359

χ^2 Table (d f = 10)

Age	R	W_1	W_2	W_3	Totals
9	2.17	2.00	5.56	2.00	11.73
10	.39	3.27	.60	2.40	6.66
11	.36	.89	6.72	6.72	14.69
12	2.86	.69	1.92	.31	5.78
Totals	5.78 (5.99)	6.85	14,80	11.43	38.86 (18.31)

same for all three wrong answers reflects the assumption that these answers are blind guesses and should behave like any other random event. It is this assumption which has achieved the greatest amount of evidence to the contrary in the exploratory studies.

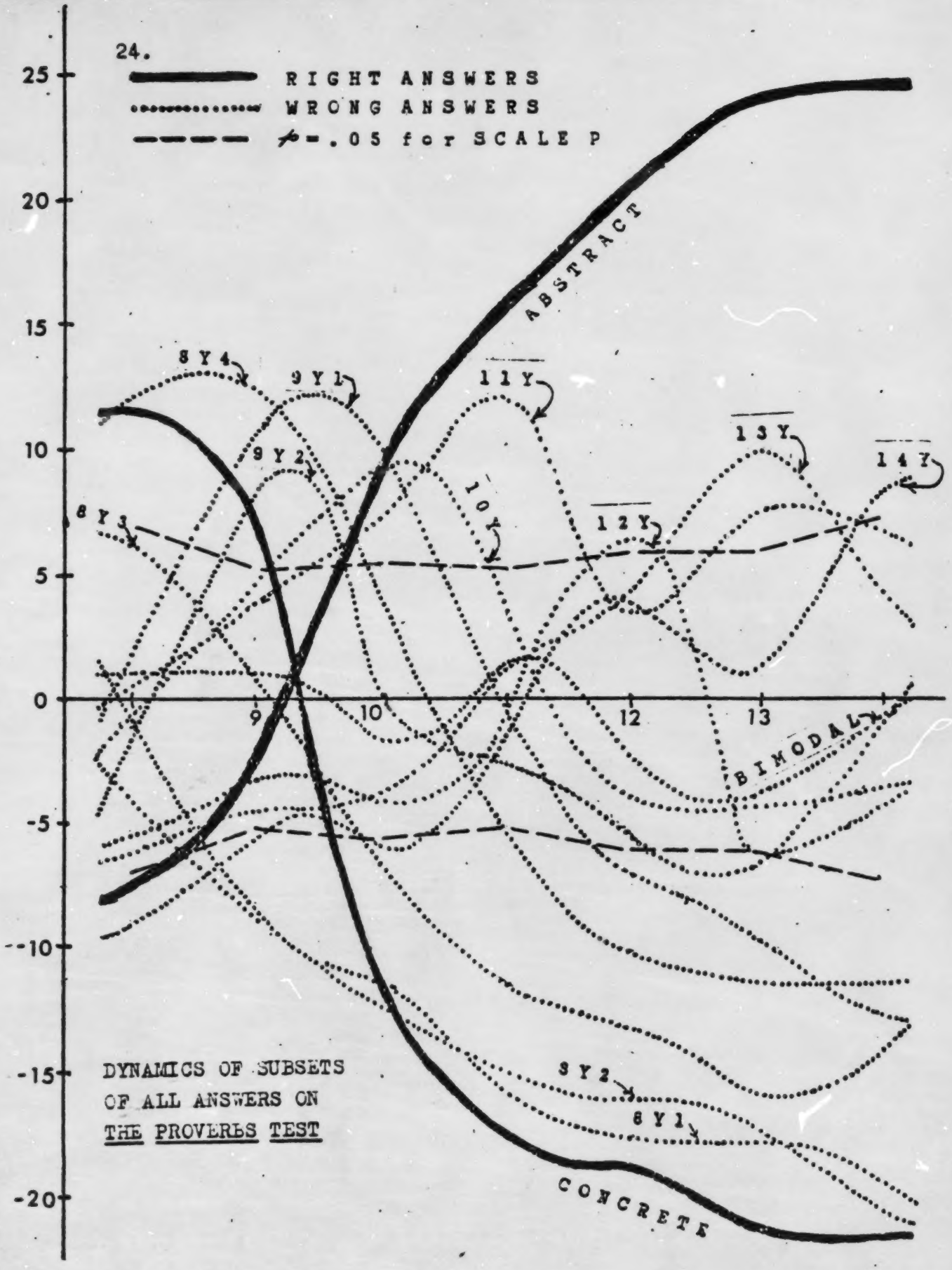
Figure 6 gives the results for Model Two, using the same format as was used for Model One in Figure 5.

INSERT FIGURE 6 ABOUT HERE

In Figure 6 the hypothetical model for the right answers is shown as a heavy line at .46 and for the wrong answers as a broken line at .18. There is an inconsistancy between the .46 used here and the .48 reported in the proportions in Figure 4. The mean proportion, correct to the fifth place is .46518. The difference is partly rounding error, and partly a result, for the .48 of the differences in group size being ignored when the average of the proportions was taken. This study is full of small differences such as this. The impact of one of them will be noted later. In any event, all the conclusions presented from the interpretation of the results should be considered in the light of a small range of error being present in the precision of all of the calculations.

To minimize this problem, since proportions were rounded to second place and frequencies to whole numbers and in both cases were adjusted to the correct totals (1.00 and the group size respectively) adjustments were made in a balanced fashion, or for the largest expected values. The balancing procedure can be inferred from the distribution of expected frequencies in Figure 5, and the use of the greatest adjustment for the largest value is found in the case of the mean proportions just discussed. In both cases the objective was to have the least possible impact. It is often the case that researchers get the main results from the computer, but do all of the "quick and easy" things by hand. It seems not so often that the appropriate precautionary note is sounded.

To facilitate interpretation of the χ^2 table, the critical values at the $p = .05$ level are indicated parenthetically below the totals. The table total has dropped dramatically with this single refinement. The improvement is visually obvious as well.


24.
RIGHT ANSWERS
WRONG ANSWERS
p = .05 for SCALE P
25
20
15
10
5
0
-5
-10
-15
-20
8 Y 4
9 Y 1
11 Y
13 Y
14 Y
9 Y 2
10 Y
12 Y
8 Y 3
9
10
12
13
ABSTRACT
BIMODAL
CONCRETE
8 Y 2
8 Y 1
DYNAMICS OF SUBSETS
OF ALL ANSWERS ON
THE PROVERBS TEST

Clearly, a large part of the departure from random among these data is found in the proportion of learners who chose the right answer, and the resulting average choice of all wrong answers. But this improvement is not good enough to have the model fit the criterion.

It is interesting to consider the impact of formula scoring upon this item. Since the average is .46 for the right answers and it is .18 for the wrong answers, only 28 per cent (.46 - .18 = .28) of all answers can be considered to be systematic events using the assumptions in Model Two. It is not surprising, in this case, that still more of the pattern can be accounted for than in this one step. Some support for current test theory is found in the large drop in the table total, but the failure of the model overall indicates that it is not sufficient.

On the basis of this latter conclusion, it is fascinating to note that, for 2 degrees of freedom, the right answer column already fits to criterion. On the basis of the total distribution this fit is spurious, since the only property it reflects is the relatively narrow oscillation of the right answer pattern in comparison to its average size. The same observation holds for the 12 year-old pattern. However, had the wrong answers been reduced to zeros, as is current practice, this analysis would have gone no further. In would seem to be obvious, without going any further, that current practice may be oversimplifying these data.

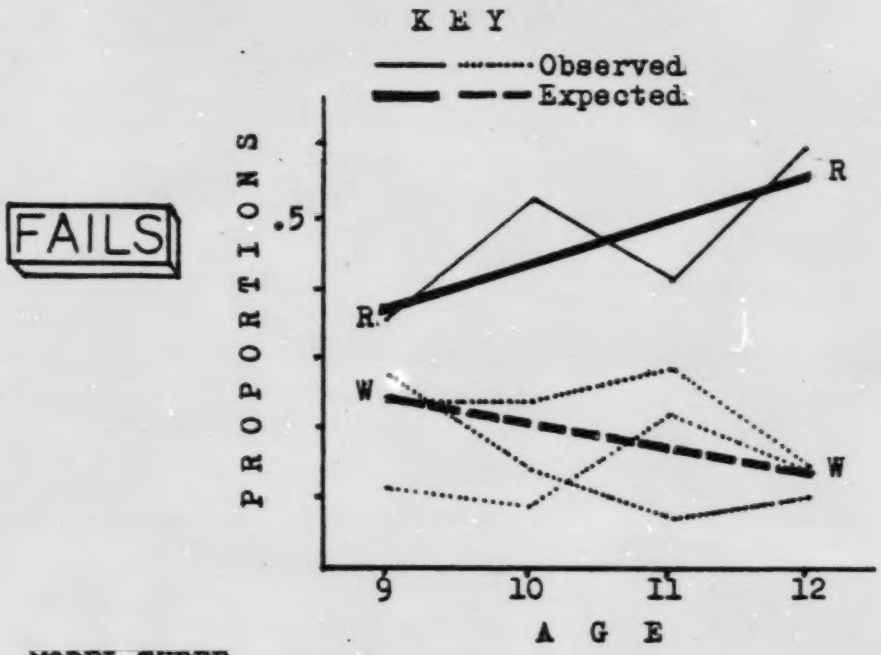
MODEL THREE: Regression Line on Right Answers.

The regression line is similar to the MULTIQUAL approach in the sense that it is also a "least squares fit". The equation for this line is derived from a procedure which minimizes the sum of the squares of the distances of points in the data set from the line and sets the sum of the directed value of these distances to zero. In this sense, the regression line is the straight-line average of the point set.

For present purposed, the line was obtained from the proportions, and adjusted slightly so that the complement of each value was evenly divisible by three in the second decimal place. The outcome of this procedure is given in the "Expected Proportions"

FIGURE 7

MODEL THREE: REGRESSION LINE ON RIGHT ANSWERS



MODEL THREE

Expected Proportions

Age	R	W_1	W_2	W_3
9	.37	.21	.21	.21
10	.43	.19	.19	.19
11	.52	.16	.16	.16
12	.58	.14	.14	.14

Observed/Expected Frequencies

Age	R	W_1	W_2	W_3	Totals
9	36_{37}	12_{21}	28_{21}	24_{21}	100
10	45_{38}	8_{16}	12_{16}	21_{16}	86
11	41_{51}	22_{16}	7_{16}	29_{16}	99
12	45_{44}	10_{10}	8_{10}	11_{10}	74
					359

χ^2 Table (d.f = 9)

Age	R	W_1	W_2	W_3	Totals
9	.03	3.86	2.33	.43	6.65
10	1.29	4.00	1.00	1.56	7.85
11	1.96	2.25	5.06	10.56	19.83
12	.02	.00	.40	.10	.52
Totals	3.30 (3.84)	10.11	8.79 (5.99)	12.65	34.85 (16.92)

section of Figure 7.

INSERT FIGURE 7 ABOUT HERE

The slope of the composite line for the wrong answers is, of course, one third of the slope for the right answers. The visual fit is also improved.

On the other hand, slight reduction of the overall value for χ^2 seems to be off-set by the loss of another degree of freedom. Contrast this observation with the improvement of fit when the right answers are considered by themselves, and the heart of the issue being addressed in this study is exposed. A reduction of these data to ones and zeros would support the reduction procedure. Such support is tautological, since it is self-defining, but is not found from the over-all results. It may be stretching the point to imply that this observation applies to all such reductions simply because it quite evidently applies to this one. If this observation is, however, a general phenomenon, then the persistance of general practice and the persistant overlooking of wrong answer phenomena could be at least partially explained.

The close fit for the 12 year-olds is also spurious. Alternative A (W_1) in fact continues its general upward trend among the 13 year olds, as shall be seen later. This particular alternative actually has its highest values among the 17 year olds. In other words, the close approach of the three wrong answer variables to each other at age 12 would seem to be an interactive, rather than a random event. As a result, the failure of the over-all model to fit these data would seem to be more founded in the data than is the apparent improvement for the right answers by themselves.

These latter comments are devived from the fact that the distribution properties of these data were well known from the "exploratory" studies of these data prior to this "confirmatory" one. These latter observations reinforce the often expressed opinion that distribution properties of a data set should be known before hypothetical question are answered from them.

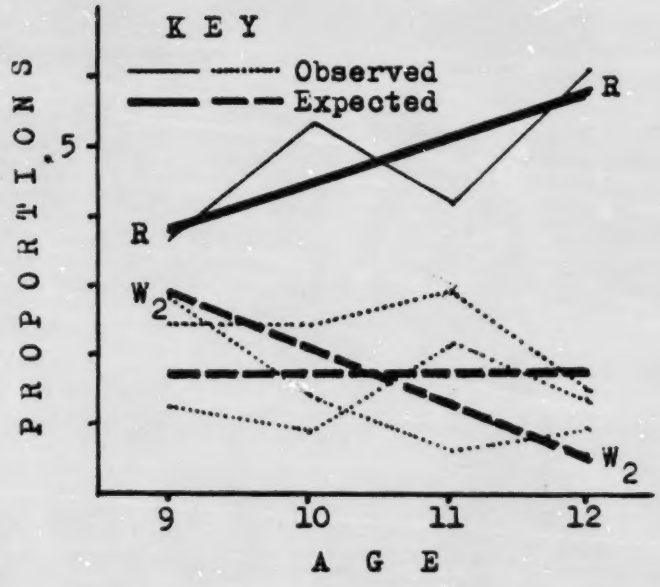
MODEL FOUR: Regression with One Wrong Answer Linearly Dependent.

Before the issue of linearity among the right answers alone

FIGURE 8

MODEL FOUR: REGRESSION
WITH ONE WRONG ANSWER
LINEARLY DEPENDENT

FAILS



MODEL FOUR

Expected Proportions

Age	R	W_1	W_2	W_3
9	.37	.18	.29	.18
10	.43	.18	.21	.18
11	.52	.18	.12	.18
12	.58	.18	.06	.18

Observed/Expected Frequencies

Age	R	W_1	W_2	W_3	Totals
9	36_{37}	12_{17}	28_{18}	24_{28}	100
10	45_{38}	8_{15}	12_{18}	21_{15}	86
11	41_{51}	22_{12}	7_{14}	29_{22}	99
12	45_{44}	10_{13}	8_{4}	11_{13}	74
					359

χ^2 Table (d f = 8)

Age	R	W_1	W_2	W_3	Totals
9	.03	1.47	5.56	.57	7.63
10	1.29	3.27	5.00	2.40	11.96
11	1.96	8.33	3.50	2.23	16.02
12	.02	.90	4.00	.31	5.23
Totals	3.30 (3.84)	13.97	18.06	5.51 (5.99)	40.84 (15.51)

is left entirely, one more issue should be addressed. This is the issue of "linear dependency." If everyone answers an item the sum of the right answers and the aggregate sum of the wrong will always equal the number of testees. This fact of linear dependency creates problems for the linear analysis of data using the General Linear Model, as is basic for current practice, for purely technical reasons. The situation for curved-line analysis of these data is less well known.

In all of the previous cases, this dependency has been equally distributed among all wrong alternatives upon the assumption that they were random events. It is also possible that only one of the wrong answers actually screens the right answer and the others act independently. If the irregularities in the pattern are assumed to be meaningless, the wrong answer which most nearly fits this property is alternative B (W_2). Figure 8 presents this relationship.

INSERT FIGURE 8 ABOUT HERE

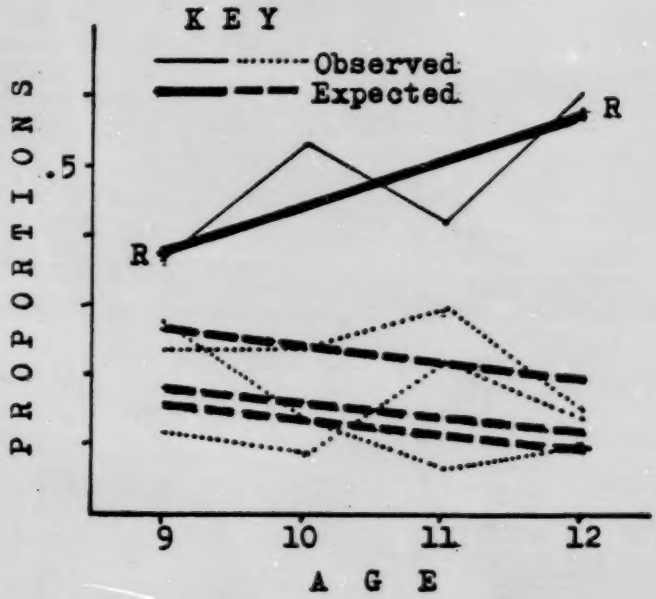
The visual fit appears as though it should be better. Actually the results deteriorate. This observation is less surprising than it would, at first, seem since an earlier exploratory study (Powell and Isbister 1974) had already shown that the data on a different test contained less of the propetries expected for linearly dependent data than is currently assumed. In this sense, Model Four is a partial replication of other findings, both of which suggest curved-line phenomena may dominate these data sets.

Another source of evidence to support the curvi-linear conclusion may be obtained by the manner in which the sizes of the cell χ^2 values change dramatically in size as parts of the expectation table are adjusted to conform to each successive model. This observation would seem to challenge the common practice , at least for the reporting of research findings, of considering only one analytical approach to a data set. It may be that not enough is known about the distribution of data sets obtained from educational studies to be able to afford the luxury of seeking a uni-dimensional conclusion. It is too easy to assume that the rejection of the null hypothesis supports only one alternative hypothesis, the one expected from

FIGURE 9

MODEL FIVE: REGRESSION WITH WRONG ANSWER PROPORTIONS

FAILS



MODEL FIVE

Expected Proportions

Age	R	W_1	W_2	W_3
9	.37	.17	.18	.27
10	.43	.15	.16	.26
11	.52	.12	.14	.22
12	.58	.10	.12	.20

Observed/Expected Frequencies

Age	R	W_1	W_2	W_3	Totals
9	36_{37}	12_{17}	28_{18}	24_{28}	100
10	45_{37}	8_{13}	12_{14}	21_{22}	86
11	41_{51}	22_{12}	7_{14}	29_{22}	99
12	45_{44}	10_{7}	8_{9}	11_{14}	74
					359

χ^2 Table (d f = 7)

Age	R	W_1	W_2	W_3	Totals
9	.03	1.47	5.56	.57	7.63
10	1.73	1.92	.29	.05	3.99
11	1.96	8.33	3.50	2.23	16.02
12	.02	1.28	.11	.64	2.05
Totals	3.74 (3.84)	13.00	9.46 (5.99)	3.49	29.69 (14.07)

the theory behind the study.

MODEL FIVE: Regression with Wrong Answer Proportions.

The analysis of Model Four completed that part of the chain of models in which wrong answers were assumed primarily to be random events. Using the same sequence of logic as before, the first property of wrong answers to assume may be meaningful would be the proportions of their selection. The regression line is retained for the right answers, simply to save the testing for fit of an additional model. Figure 9 gives the results of this analysis.

INSERT FIGURE 9 ABOUT HERE

Once again there is a reduction in the size of the over-all total, although this may also be largely offset by the loss of another degree of freedom.

There is also a slight loss for the right answers because of the rounding problems already mentioned. Carrying the calcutations to a greater number of digits than two might clarify whether or not a slight advantage might be gained from this approach. In this middle age range, two other studies (Powell 1979, 1980) have shown that the potential contribution of wrong answers may be about equal to that of the right ones. The value of wrong answers seems to increase with age. From these latter observations, the value of further refinement, at least for present purposes may be open to question. This issue will be returned to later.

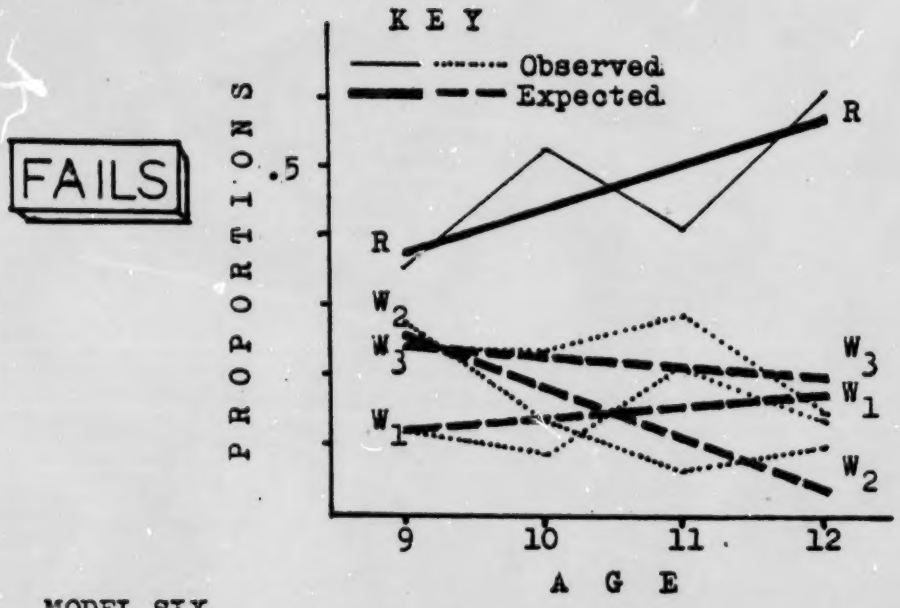
In any event, the sizes of the cell χ^2 s seems to be dropping, suggesting that the dynamics of this chain is moving in the right direction.

MODEL SIX: Regression lines on all Variables.

This model is the last of the strictly linear models which will be tested. There are several possible approaches to the setting up of this model. In this case, the proportions of selection were regressed, and the resulting lines were adjusted to total the appropriate 1.00 which has been maintained throughout. Slight distortions will be the inevitalbe result of this procedure.

FIGURE 10

MODEL SIX: REGRESSION LINES ON ALL VARIABLES



MODEL SIX

Expected Proportions

Age	R	W_1	W_2	W_3
9	.37	.12	.26	.25
10	.43	.14	.19	.24
11	.52	.15	.11	.22
12	.58	.17	.04	.21

Observed/Expected Frequencies

Age	R	W_1	W_2	W_3	Totals
9	36_{37}	12_{12}	24_{26}	24_{25}	100
10	45_{37}	8_{12}	12_{16}	21_{21}	86
11	41_{51}	22_{15}	7_{11}	29_{22}	99
12	45_{44}	10_{13}	8_{3}	11_{14}	74
					359

χ^2 Table (d f = 5)

Age	R	W_1	W_2	W_3	Totals
9	.03	.00	.15	.04	.22
10	1.73	1.33	1.00	.00	4.06
11	1.96	3.13	1.45	2.23	8.77
12	.02	.90	8.33	.81	10.06
Totals	3.74 (3.84)	5.36	10.93	3.08	23.11 (11.07)

Figure 10 gives the outcomes of this analysis.

INSERT FIGURE 10 ABOUT HERE

Once again improvement is evident, and once again, the amount of improvement is small in relation to the loss of another degree of freedom. The observed total χ^2 is still somewhat more than twice the critical value for the p = .05 level.

In a methodological study such as this one, a number of options are open for the design of models to fit. In the one chosen, the proportions were used in the regression equations. This procedure removes the impact of different group sizes, but distorts the actual relationships among the observed frequencies. The regressed proportions were then rescaled so that the total proportions for each age level were restored to 1.00. This procedure made it possible to have the total expected frequencies match the group sizes, but made the regression lines not quite straight. For the right answers for instance, where the least impact upon χ^2 will occur, the progressive changes are +.06, +.09 and +.06 where +.07 would have given a straight line.

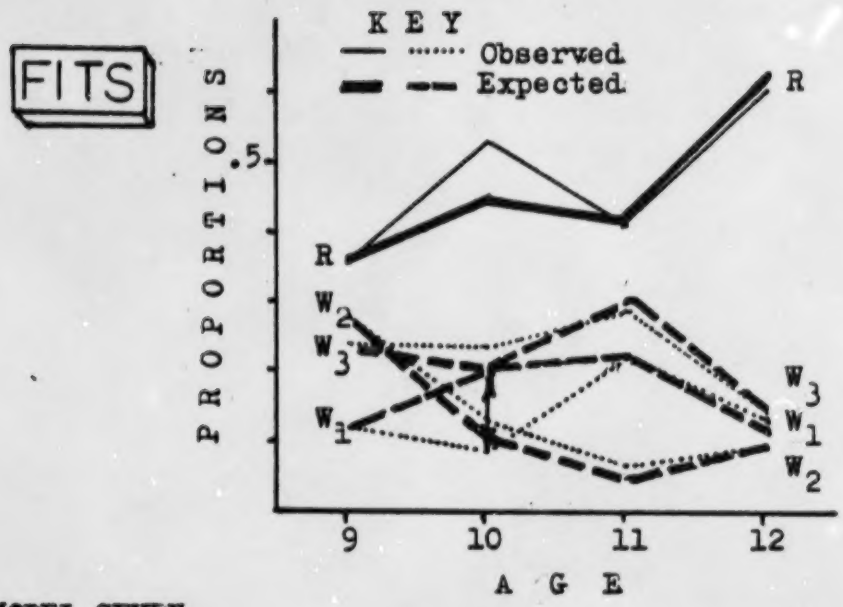
It would have been possible to have regressed the observed frequencies instead, with or without adjustment to restore group size. If the group size is not restored, either through the proportions, or the frequencies themselves, then the sum of the expected frequencies will not equal the total group size. This violation of the assumptions behind χ^2 would be expected to become increasisngly serious, the greater the departure. Using greater numerical precision and repeated regression are other possibilities.

Such refinements are easily possible for someone with the time and an interactive computer facility, and improvement in the overall total could easily be found. Of critical importance is the relative size for the expected value for W_2 at age level 12, where more that one third of the overall total is generated. The dynamics of the observed pattern requires this value to be large, but how large is a matter of the manipulation procedure used.

This discussion illustrates very well the importance of assumptions in the conducting of research. The central issue behind this

FIGURE 11

MODEL SEVEN: MULTIQUAL
FIT OF POLYNOMIALS
TO THE THIRD DEGREE



MODEL SEVEN

Expected Proportions

Age	R	W_1	W_2	W_3
9	.36	.12	.28	.24
10	.45	.21	.13	.21
11	.42	.22	.06	.30
12	.62	.13	.10	.15

Observed/Expected Frequencies

Age	R	W_1	W_2	W_3	Total
9	$36/36$	$12/12$	$24/28$	$24/24$	100
10	$45/39$	$8/18$	$12/11$	$21/18$	86
11	$41/41$	$22/22$	$7/6$	$29/30$	99
12	$45/46$	$10/10$	$8/7$	$11/11$	74
					359

χ^2 Table (d f = 4)

Age	R	W_1	W_2	W_3	Totals
9	.00	.00	.57	.00	.57
10	.92	5.56	.09	.50	7.07
11	.00	.00	.17	.03	.20
12	.02	.00	.14	.00	.16
Totals	.94	5.56	.97	.53	8.00 (9.49)

entire study is the importance of making appropriate assumptions for the nature of the behavior and of the data set being studies. The procedure used for Model Six can be defended upon the basis that it is conservative. and is consistant with the procedures used in the other models.

In any event, none of the linear models seem to have fitted this data set. Even though other models are possible, in general they are refinements of the ones tested here. It is reasonable to expect that such refinements will not have any major effect upon this general conclusion. The most important step in this chain is the next one. To demonstrate conclusively the invalivity of the use of dichotomization upon an answer set for this particular item it is necessary for the MULTIQUAL Model to fit this data set, and for the reduction in the overall total to be convincingly large.

MODEL SEVEN: MULTIQUAL Fit of Polynomials to the Third Degree.

As is clearly evident in Figure 11, both of these requirements have been met.

INSERT FIGURE 11 ABOUT HERE

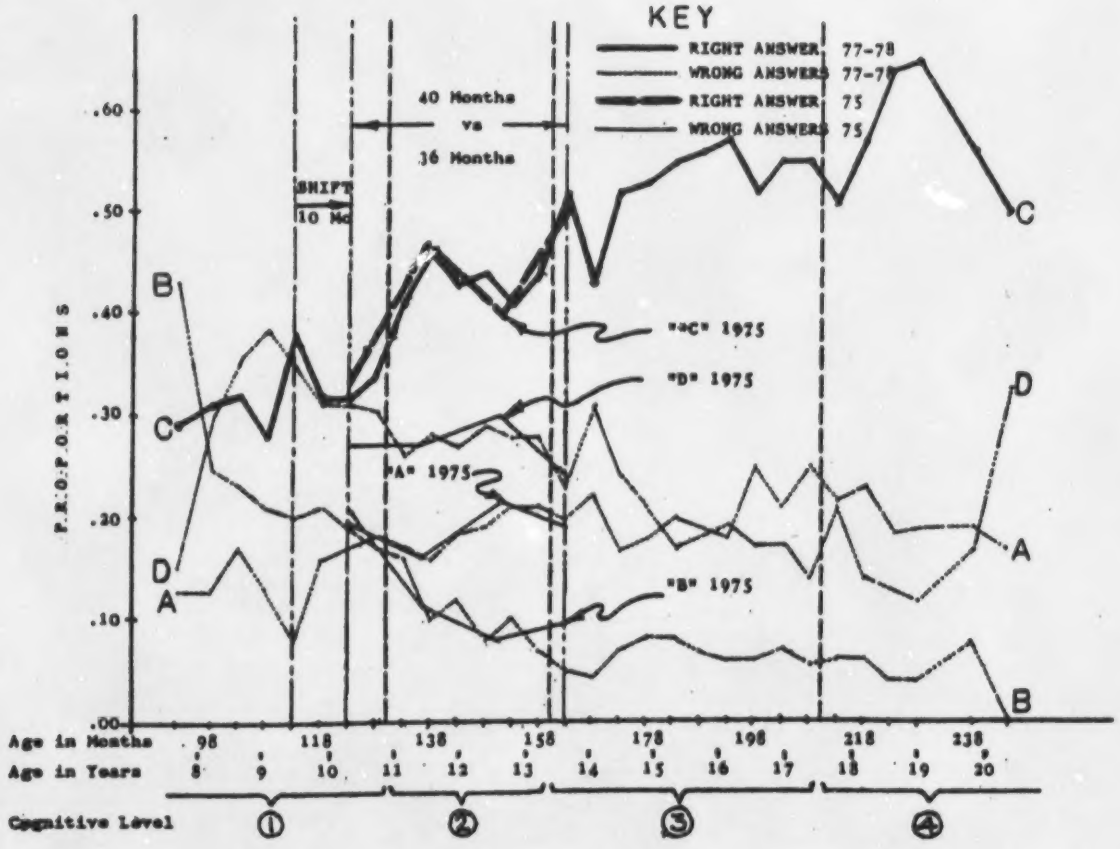
Not only is this the only model which fits the data set, but also the overall total is about one third of its previous value.

The MULTIQUAL approach seems to have fitted cubic functions to the right answers and to W_3 (alternative D), and parabolas to the other two alternatives. The arrow at age 10 for W_1 shows the point where the greatest departure from expected and observed occured. A cubic function would have fitted this set better. Almost 70 per cent of the χ^2 total is found in this one departure.

For most purposes, and for short ranges of variables, polynomials will give good approximations of more complex curves. It was already known (Yu 1977) that the best fitting curves for a randomly selected set of 20 of the 160 variables in this test, when the full available range was used, were negative exponentials. Three points in a sequence are usually not considered enough to distinguish between a straight and a curved line, or among curves. Hence a four point sequence was used in this study.

FIGURE 12

ATTEMPT AT REPLICATION
OF THE PATTERN



If this study had been the first in this series, little would have been demonstrated by it other than the likelihood that curved-line phenomena were present in this item, and perhaps in this test. However, this particular study comes at the end of a series of studies involving more than one test, and several different populations, all of which showed indications that curved-line phenomena might be present. It is in this latter sense that this study is "confirmatory". Curved-line phenomena seems to be much more common than had been suspected. In fact, current test theory proceeds upon the assumption that such phenomena are NOT present among test data in any meaningful amounts.

It is not sufficent, for the complete invalidation of current test theory to base such a decision upon this study alone. The only issue addressed here was the validity of the dichotomization of answers within one item. This study showa that this much, and only this much, of current test theory may be invalid for at least some tests.

This potential invalidity could be strengthened if the MULTI-UAL Model and/or the original data set was shown to replicate.

REPLICATION STUDY

The same test was given to eight elementary schools in the same community on a repetition basis 17 and 22 months later than the data collection for the present one. The resulting pattern of distribution using 5 month intervals was compared to the pattern found here. Figure 12 gives the results of this comparison.

INSERT FIGURE 12 ABOUT HERE

The basic data shown in Figure 12 is the aggregate of the October 1977 and March 1978 administrations using a 5 month age band on the age of the Subjects so that noone would be counted twice.

A reasonably good fit was found if transformations of the data set are permissible. Using the sign test to equate the general direction of change, 23 of the 32 possible comparisons were found to be going in the same direction as the 16 point set provided. This means that these two patterns are significantly similar at the

p = .05 level. (z = 2.48)

No theoretical accounts for the pattern were made when the visual overlay was attempted. Any changes which were made, were applied in an equivalent manner for all variables. The lower limit for the pattern was raised by 10 months and the upper by 14. It should be noted that the second administration was 17 and 22 months after the first. In addition, six new schools and all of the children, rather than a sample of the two earlier schools which were in the new group were included. The possibility that the same choice pattern was made by the same group must be discounted.

The other changes which were made were based upon the fact that the selection ratios were different for the new group. Hence changes in vertival position of the points was made, without changing the general shape of the pattern or the constraint that the new pattern needed to add to 1.00 within each age group. The original group contained 359 Subjects and the new one 2292 Subjects.

For whatever reason, there is a surprising similarity between these two groups on thia section of the age range for this item. Finding such a high degree of similarity for a complex curved-line pattern in the context presented here is sufficiently unlikely that this event is probably conclusive, even with the transformations needed to obtain a fit.

The only choice seems to be to assume that for some tests at least, the dichotomization of answers into "right" and "wrong" categories would seem to be invalid, both upon psychological and upon statistical grounds.

This finding does NOT invalidate the use of aggregation of "right" answers into a total correct score, or the assumption that the total-correct scores can be used as an interval scale. It took two other studies (Powell 1979, 1980) to achieve this latter disconfirmation. On the other hand, the amount of curved-line phenomena found here puts the entire current test theory upon shakey ground.

Some procedure designed to organize the sequence of models investigated here into a single pattern would assist the summation.

A PATTERN OF MODELS

The objective of this section is to attempt to examine several possible ways to find a continuity, other than the logical one used in their development, within the sequence of results which emerged. Table 2 gives these results.

INSERT TABLE 2 ABOUT HERE

In Table 2 the results from three different summary attempts are recorded. The first of these is an adaptation of the use of the Contingency coefficient (C). Since this figure is similar to a correlarion, but this study is looking for a reverse relationship, the simple expedient of using the complement (1 - C) was employed. The square of this value, if the procedure is appropriate should represent the amount of variance accounted for. The values in the L^2 column are too high in relation to the experience and from the exploratory studies to be valid for this purpose.

The second procedure, of dividing the observed total χ^2 by its critical value at the .05 level shows the deterioration of Model Four, and the gain for Model Seven, but does not discriminate among the others. Since the χ^2 distribution becomes more peaked as the number of degrees of freedom becomes smaller, the tail, which contains the critical values becomes flatter. As a result, the ratio of one value to another in this latter range increases at the same pair of probibility levels. As a result, keeping the same relative ratio for decreasing degrees of freedom implies slight improvement. How much improvement, however, can not currently be determined using this procedure.

The third approach involves assuming that the total χ^2 for departure from random (Model One) represents the best available estimate for the total available variance. A procedure to determine a coefficient of fit (K) was developed from this assumption. The use of the square root was suggested for the ratio in a desire to approximate a correlation coefficient. The formulae for all three procedures are given below Table 2.

This last procedure produced the most agreeable seeming results.

TABLE 2

ATTEMPTS TO FIND A
SUMMARIZING PATTERN

Model Number (n)	Obtained value for χ^2	Critical value for $\chi^2_{.05}$	J	C	L	L^2	K	K^2
1	129.46	19.98	5.18	.51	.41	.17	.00	.00
2	38.86	18.31	2.12	.31	.64	.41	.45	.20
3	34.85	16.92	2.06	.30	.66	.44	.48	.23
4	40.84	15.51	2.59	.32	.63	.40	.43	.18
5	29.69	14.07	2.11	.28	.68	.46	.52	.27
6	23.11	11.07	2.09	.25	.71	.50	.58	.33
7	8.00	9.49	.84	.15	.83	.69	.75	.56

FORMULAE:

(1) $$J_n = \frac{\chi_n^2}{\chi^2_{n(.05)}}$$

(2) $$C_n = \sqrt{\frac{\chi_n^2}{N + \chi_n^2}}$$ Where N is the group size (359)

(3) $$L_n = 1 - \frac{C_n}{C_{max}}$$ Where $C_{max} = \sqrt{\frac{k - 1}{k}}$ and k is the smaller of the dimensions of the contingency matrix.

(4) $$K_n = \frac{\sqrt{\chi_1^2} - \sqrt{\chi_n^2}}{\sqrt{\chi_1^2}}$$

The sequence suggests a step function with regression lines somewhat better than proportions. The steps are; 1) a linear approach using right answers alone, 2) a linear combination of right and wrong answers. and 3) a curvi-linear pattern. This interpretation seems to be intuitively satisfying. The mathematical properties of this K distribution are not yet known.

If the K is treated like a correlation coefficient and squared for variance estimates, the results seem to conform to the implications from the exploratory studies, with about one quarter of the variance available from the regressed right answers alone, about one third from right and wrong together and more than half from the curved-line pattern. These latter figures are derived from only one item using a procedure of uncertain quality. On the other hand, they conform rather closely to the levels which the exploratory studies suggest may be available. These latter observations further support the reasonability of the curved-line approach (or the multiple-category approach) to item data.

From this vantage-point, it is now appropriate to summarize, to draw conclusions, and to develop some implications from this study.

CONCLUSIONS AND IMPLICATIONS

This present study is the first in a series of three studies which are designed to combine confirmation with replication following several exploratory studies. These studies each address the validity or lack thereof of one of the three basic assumptions behind current test theory. The purpose of this study was to address the assumption that it is valid to dichotomize answer selection into wrong answers and right answers by examining the distribution properties of the wrong answers.

The general conclusion reached in this study seems to be that this dichotomization is probably widely invalid. The broad generalizability if this conclusion arises from the procedural characteristics of this study which involved fitting a number of models to a distribution in support of or attack upon a number of findings from exploratory studies. The findings of the exploratory studies were confirmed, to the disconfirmation of current practice.

Within the framework of the philosophy of science, this finding in conjunction with the other two studies, which seem to disconfirm the other two assumptions probably constitutes a breakthrough in the scientific measurement of educational achievement.

At least, these observations display several properties which would be characteristic of a breakthrough. First, the current approach seems to have been content with the reduction of all wrong answers to zero without ever validating this approach. Evidence from within this study would suggest that this course of action could lead to a spurious apparent support for the practice because it eliminated the data needed to disconfirm the practice from further analysis.

Second, this study has shown that not only do the wrong answers seem to be meaningful, but also so are the irregularities in the selection proportions which be removed when regression lines are fitted. As a result current practice may have led to the loss of two sources for meaningful information about learning processes and learner status. Thus these practices, which are in broad general use would seem to have the potential of severely over-simplifying the distribution properties of the available data in a data set from multiple-choice tests.

Third, there is evidence from both the exploratory studies and from this one that a substantial increase in the accountable variance may be available from the development of additional if not alternate practices in test analysis.

It was suggested that literature in this area has generally been sparce and has often been contradictory. This situation created the need for exploratory studies to investigate the distribution properties of wrong answers. Based upon these explorations several important conclusions were reached. There was evidence to suggest that many of the answers considered to be wrong were, in fact, correct. These answers seemed to arise from legitimate alternative interpretations of the items in queation. Instead of reflecting badly upon these items, however, it was found that these alternatives could be identified from the choice of answer more than 50 per cent of the time. As a result, the conclusion was drawn that the predetermined binary keying of such tests probably represented the invalid superimposition of a deterministic model for answer selection upon a

probabilistic behavioral event. As such, current prectice would seem to represent yet another example of the utilization of the determinism fallacy, an interpretive error which has been recognized in many other areas of knowledge over the past 100 years.

Three developmental phases seem to be present in the age range of 8 to 14+ (and five for 7 to 20+) and the strength of influence upon answer selection begins with this developmental aspect being the most powerful as reflected in the point-of-view of the learner. This aspect may be related to the way in which the learner derives meaning from his or her environment at these several different levels.

The second most powerful influence seemed to be the solution strategy selected, with the level of information recalled coming in a distant third.

These findings, as supported by the curved-line fit which they implied, have left current practice with two serious dilemmas. First, how does an evaluator distinguish, on a "right-wrong" basis those learners who get an answer scored as correct when one of them may have memorized, another understood, a third made mutually cancelling errors, and so on. The second dilemma involves using a predetermined key for answers which diverge from intention for the test constructor, but which are internally consistant and logically correct.

Thus, the test of a chain of models, only the curved-line one fitting, and having these results replicate, along with the inability of the "know-guess" model to deal with the above dilemmas has served to cast very serious doubt upon the validity of current practice.

The use of "zero-one" scoring for multiple-choice tests should, therefore, be henceforth considered to be invalid as the only basis for assessing learner status for all tests unless the validity has been demonstrated for that test using the entire data set, and not merely the right answers.

Implications.

The above statement does not preclude the use of total-correct scores, which from consistant experience seems to account for about one quarter of the variance. In fact, there is evidence to support the use of the "know-guess" approach for tests which measure recognition, recall and other fundamental lower order skills.

The issue instead seems to be that answer selection patterns should also be used for the insights they provide into the nature of the reasoning and the nature of the errors being perpetrated by the learner. The reasons for the selection of an answer seem to be as least as important as the answer itself.

Since the replication did not repeat at the original age level it can not be entirely a developmental property. Since it did not repeat at the age level appropriate for the same group as was in the original study, a population mix hypothesis does not fit these observations either. The change in proportions with the adding of city-center children, however, does support the population mix hypothesis. Probably, both are involved. In this case, it may be possible to identify quite a few learner characteristics from the nature of and the reasons for answer selection.

Clearly, intellectual development appears to be a much more complex process than is implied within the framework of the "know-guess" hypothesis. For this reason, an alternative model for learning will be proposed. This model proposes learning to be arrayed upon two dimensions. The first involves knowledge and its use as related to the practical problems of the physical environment. The second dimension relates to knowledge and its use as related to the various ways of representing real and imaginary environments in symbolic terms. A two-by-two schematic of this model is given in Figure 13.

FIGURE 13

HYPOTHETICAL LEARNING MODEL

Dealing symbolically with the Environment \ Dealing physically with the Environment	
KNOWLEDGE	APPLICATION AND TRANSFER
DEDUCTIVE AND INDUCTIVE REASONING	CREATIVITY AND INTUITION

This hypothetical model has been designed to accommodate the results thus far obtained from "wrong answer" studies. It has long been known, and has been further supported here that the use of right answers to produce a total-correct score seems to explain about one quarter of the variance. As such in this model it is represented by the basic position in the top left. The intended implication is that this 25 percent of human learning may be fundamental both to application and to symbolic learning. Application is inferred here to be derived from the meaning formulation approach being used by the learner. It has been shown in these studies to be developmental in nature and to be available from the "wrong" answers. The reasoning procedures being employed also seem to be available from the "wromg" answers and seem to be orthogonal to the meaning approach, and to be less potent.

Thus, a good deal of information about two more cells in this model seem to be available to the observer who is prepared to go beyond total-correct scores (or singular consideration of right answers in whatever manner, such as Rasch scaling). There is also some indication for those who have "over-read" the questions that particular answer selection may also signal aspects of the fourth cell.

From this approach, it may be inferred that not only do these findings help to make legitimate types of learning which go beyond the lower level skills such as recognition and recall, but also seems to point a direction toward procedures which could make these higher order learnings more precisely measureable. It seems that this approach may have create a situation wherein more variance can be explained, enabling greater precision in measurement and the measurement of a greater variety of outcomes than are currently available.

Another aspect of this work arises from the implication that the learner's meaning systems can be observed from "wrong answers." This portends much more effective observation, but it also implies the possibility that the conclusions from this study may have a more broad behavioral applicability than merely the taking of a multiple-choice test. Departures from expectation for all learner responses, M/C tests, free-responses tests and even answers to

questions in the classroom and to written assignments may all show these same properties.

If this latter inference is supported from futher research, then this challenge to current procedure may have opened the door upon an observation technique which could lead to the improvement of understanding of learning processes, to the more precise identification of current learner status, to the more precise tracking of learner progress, and perhaps even to an improvement in current ability to assess and to predict the impact of specific interventions upon precisely defined sub-populations. All of these possibilities suggest the potential for the improvement of current educational practice, not just current measurement theory.

It should be pointed out that most of these inferences have been derived from the selection patterns within an item. This present study did not consider relationships among alternatives between items, nor relationships among alternatives within an item as these change within a person over time, nor has any consideration been given to repationships among alternatives between two or more tests. Evidence from other studies in this series seem to suggest that each if these possibilities may also be able to contribute uniquely to the explained variance.

It may even be possible, in the future to give a short battery of tests, perhaps less than 100 items in total, and because of the information properties built deliberately into each alternative, be able to obtain a complete general educatioal diagnosis and placement statement from the resulting answer vector for each learner.

There is another possibility which needs consideration. Current testing practice seems to extend the tendancy to consider only the "Right " answers to other forms of assessment than only the multiple-choice test. The curved-line phenomena clearly indicated here may lead to unstable "right answer" results for tests which demand (or perhaps even assignments which demand) more than recall or recitation from the learners. Teachers in fairness to their responsibility may feel that their prediction of success should be reinforced by outcomes. If right answers can give meaningful results only for memorization, then current testing practice may be forcing the memorization mode of learning upon current classroom practice.

Furthermore, since right answers achieved through memorization can not be distinguished from right answers achieved through understanding when the observation is confined to the right answer set, less learning than is commonly assumed may be occurring. With the possibility that deficits in understanding would be expected to accumulate over the years, current practice in testing may be curtailing achievement far below the potentials of most learners.

This implications section has presented a large number of researchable hypotheses. It is impossible to imagine where an emergent technology such as this might lead. The prospects of achieving two or more times as much explained variance as is now common, itself portends substantial advances in knowledge about the phenomena being observed. With increased knowledge comes the potential of increased ability to control. In that direction lies much danger unless the ability to control is shared by teaching learners how to control themselves. This last comment returns the discussion to the philosophical level from which it began. The lesson of history has been that it is not power but the abuse of power which has created human misery. Whatever our persuation, the knowledge of the nature of existance is there for someone to find and to use or to abuse. In the end, for better or for worse some human somewhere must make a decision. It is the learning how to make wise decisions which is our ultimate hope. If wisdom comes from learning how to detect and how to correct errors then attention to the "wrong" answers in the instructional process may be long overdue.

REFERENCES

Bock, R. Darrell. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 37. (1). March 1972. 29 - 51.

Bock, R. Darrell. MULTIQUAL: Log-linear analysis of nominal or ordinal qualitative data by the method of maximum likelihood. Chicago, National Educational Resources, 1973.

Fouldes, G. A. and Forbes, A. R. Advanced Progressive Matrices: Sets I and II (manual). London. H. K. Lewis. 1965.

Glasser, Barney G. and Strauss, Anselm L. The Discovery of Grounded Theory: Strategies for Qualitiative Research. Chicago. Aldine, 1967.

Gorham, Don R. The Proverbs Test. Missoula Montana. Psychological Test Specialists, 1956.

Hakstian, A. Ralph and Kansup, Wanlop. A comparison of several methosd of assessing partial knowledge in multiple-choice tests: II. Testing procedures. Journal of Educational Measurement. 12 (4), 1975, 231-254.

Hoffman, Banesh, The Tyranny of Testing. N. Y., Cowell-Collier Press. 1962.

Jacobs, P. I. and Vandeventer, M. Information from Wrong Responses. Research Bulletin RB-68-25. Princeton, Educational Testing Service, June 1968.

Kansup, Wanlop and Hakstian, A. Ralph. A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. Journal of Educational Measurement. 12 (4). 1975. 219-230.

Powell. J. C. The interpretation of wrong answers from a multiple choice test. Educational and Psychological Measurement. 28 (2). 1968. 403-412.

Powell. J. C. Achievement information from Wrong Answers. (Short title). Unpublished Ph. D. Dissertation. Edmonton. Alberta. University of Alberta, May 1970.

Powell, J. C. Evidence for a phase and stage developmental sequence derived from response patterns on multiple choice tests. Paper presented to the Annual Convention of the American Psychological Association. Washington, D. C.. Sept. 1976, 62pp.

Powell, J. C. The developmental sequence of cognition as revealed by wrong answers. Alberta Journal of Educational Research. 23 (1), 1977, 43-51.

Powell, J. C. Can developmental status information be obtained from wrong answers? Paper presented to the National Council on Measurement in Education. San Fransisco. April, 1979.

Powell. J. C. Patterns of change among answer selection. Paper presented to the National Council on Measurement in Education, Boston, April. 1980.

Powell, J. C. and Isbister, Alvin G. A comparison between right and wrong answers on a multiple-choice test. Educational and Psychological Measurement, 34 (3). 1974. 499-509.

Shuford, E. H., Jr., Albert, A. and Massengill, H. Admissable probability measurement procedures. Psychometrika, 31, 1966, 125-145.

Yu, Kenneth, Developmental patterns on multiple choice rests of elementary students. Unpublished major paper. Department of Mathematics, University of Windsor, June 1977.